Deep Learning for Supporting Ocean Data Quality Control

Serdar Demirel, Alfred-Wegener-Institute (Germany), serdar.demirel@awi.de
Sebastian Mieruch, Alfred-Wegener-Institute (Germany), sebastian.mieruch@awi.de
Reiner Schlitzer, Alfred-Wegener-Institute (Germany), reiner.schlitzer@awi.de

The frame and quality control

This project is part of a large European program, the SeaDataNet (2004-2016) and SeaDataCloud (2016-2020) initiatives, which have the aim to provide quality controlled ocean data via web services. Since 2004, more than 100 European data centers have been included, which provide rich data and meta data collections of several variables like temperature, salinity, oxygen, nitrate etc. for the global ocean with focus on the European seas.

In order to provide oceanographic data for research purposes, each sample of the measured variables (temperature, salinity, oxygen, nitrate etc.) is flagged by data providers and SeaDataNet ocean experts manually and/or visually. Additionally, the ocean experts have set up semi-automated workflows for the QC that consist of classical range and distribution checks (e.g. Simoncelli et al., 2018). Experts also use the Ocean Data View Software (Schlitzer, 2002) that allows them to label data with quality flags (QF). Every single ocean profile is scanned, while searching for outliers, anomalies and erroneous data. Since the size of the dataset is enormous (ca. 9 million profiles) and it is expected to grow significantly in the future, the automation/semi-automation of the QC is a necessity for the ocean science community.

Data and deep learning approach

Thus, the main motivation of our project is to assist QC experts by supporting the automation/semi-automation of the QC procedures. For this purpose, the state-of-the-art methods from the field of Artificial Intelligence (AI), specifically deep learning algorithms, are used. A binary classification problem is considered which is aiming at detecting outliers on measured temperature data in millions of oceanographic samples. For this reason, a Multilayer Perceptron (MLP) neural network, which is a class of feedforward artificial neural network (ANN), is designed which we name *Salacia* according to the roman goddess of sea water. To this end, in the framework of the EU SeaDataNet infrastructure, *Salacia* is trained with the already quality controlled "Mediterranean Sea - Temperature and salinity observation collection V2" (Simoncelli et al., 2015), where we use only data east of Gibraltar. For efficient training purposes, we sub-sampled the dataset to include only profiles with one to 100 *bad* flagged samples (and the rest *good*), to consider surely all bad flagged data. Finally, we removed the gross outliers by range filters (e.g. temperatures above 40 °C), and came up with 141,295 profiles containing 2,080,698 temperature samples. The most important aspects in machine learning are (**i**) the input features for the algorithm, i.e. the information about our data that is fed into the model, (**ii**) the separation of the dataset, and (**iii**) the architecture of the neural network.

- *Salacia* uses the most basic and informative features, which are also available for the QC experts, which are listed in the following for each sample: Depth, Temperature, Longitude, Latitude, Season (Month), Temperature gradient (change of temperature with depth), Temperature gradient from the sample above, and Temperature gradient from the sample below.
- The dataset is divided into four parts: Training data (55 %): to be used to train the network, Validation data (15 %): to tune the data to avoid under- and overfitting, Testing data (10 %): to tune the classification thresholds, Control data (20 %): to assess the skill of the model.
- A fully connected network of 3 hidden layers and 32 nodes each has been chosen.

Results and next steps

Similar to Simoncelli et al. (2018), the Mediterranean Sea has been divided into 16 Regions to evaluate the skill of Salacia on a regional scale. Figure 1 shows an example of the skill assessment for a region between Italy and Libya.

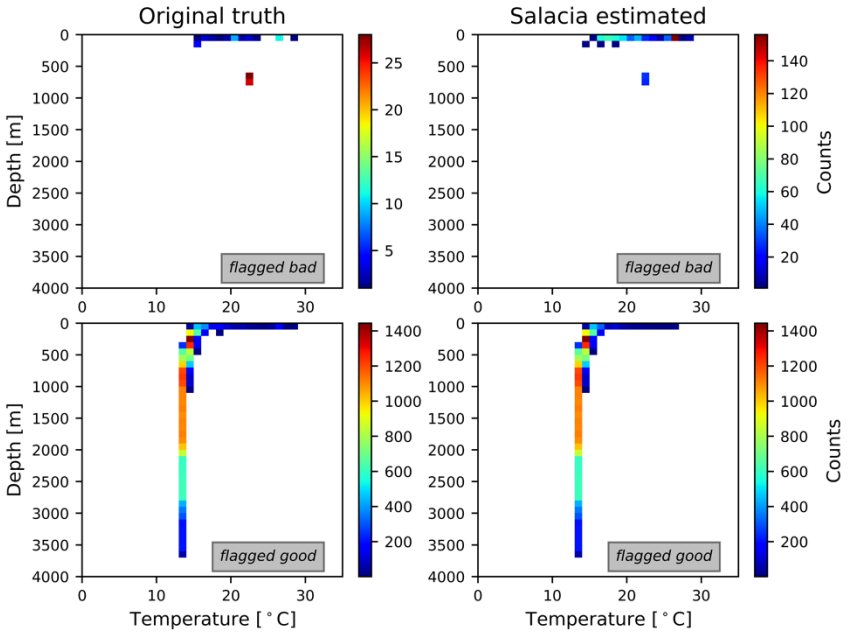

Figure 1: Temperature / Depth plots for truth (left) vs. Salacia (right). *Salacia* predicted with accuracy larger than 90 % for both *good* (green) and *bad* (red) flags.

The left side of Figure 1 shows the reference temperature measurements, flagged *bad* (top) and *good* (bottom) by the human QC experts. The right side shows the estimated flags by our algorithm *Salacia*. It is important to note that this evaluation data is "unknown" to the algorithm, i.e. it has not been used during the training process. Our evaluation reveals that *Salacia* is too sensitive in classifying data as *bad*. However, among the 646 *bad* classified samples by *Salacia*, 85 from 94 true *bad* have been found correctly (90.42 %). Regarding the *good* flagged samples, *Salacia* has found 32898 out of 33459 correctly (98.32 %). In general, we have found that the algorithm reaches high accuracies larger than 90 % in identifying *good* or *bad* data in 11 of 16 regions of the Mediterranean Sea, and for the rest of the regions, the accuracy values are oscillating between 90 and 60 % for *good* and *bad* data. Thus, Salacia could be especially useful and helpful for the QC experts in these particular skillful regions. However, it would be recommended that the QC experts concentrate on using only the *Salacia bad* flags as a guidance and accept the *good* flags. This leads to checking only ca. 10 % of the data. Now, the crucial question is if *Salacia* can be an assistant for the QC experts by giving useful hints to potentially *bad* data on the small scale. This has to be evaluated together with the QC experts.

References

Schlitzer, R. (2002). Interactive analysis and visualization of geoscience data with ocean data view. Computers & geosciences28, 1211–1218

Simoncelli, S., Coatanoan, C., and Myroshnychenko, V. (2018). Seadatacloud temperature and salinity historical data collection for the mediterranean sea (version 1). product information document (pidoc)

Simona Simoncelli, Dick Schaap, Reiner Schlitzer (2015). Mediterranean Sea - Temperature and salinity observation collection V2. https://doi.org/10.12770/8c3bd19b-9687-429c-a232-48b10478581c